# Implicatures and the recognition of speaker intentions in a game theoretic model

Anton Benz
Center for General Linguistics, Berlin

**Abstract**

In this paper we study the role of the recognition of the speaker's intentions in utterance interpretation. We are interested in the question whether the recognition has to be explicit and whether it precedes or follows interpretation. We present a game theoretic model of pragmatic interpretation and study it in the context of iterated best response models as these models make the interlocutor's reasoning about each other explicit. We claim that for normal communication backward induction is sufficient for interpretation, and hence that the recognition of the speaker's intentions follows rather than precedes interpretation. Relevance implicatures are generally considered to be part of the communicated meaning. By definition, it seems that the hearer has to take the speaker into account when calculating these implicatures. We will therefore pay special attention to them.

## 1 Introduction

Communication is an intentional activity, and it is widely held that the recognition of the speaker's intention is a necessary part of successful interpretation. In this paper, we discuss the validity of this claim with respect to particularise implicatures in a game theoretic model of communication (Benz and van Rooij, 2007).

According to (Grice, 1957), linguistic communication is characterised by the fact that the recognition of the speaker's intention is a necessary step in bringing about the speaker's intended effect in his audience. The intended effect may be a belief, or an obligation for the speaker or the hearer. If the intended effect is a belief, it can be naturally called the *speaker's meaning* of an utterance. Grice proposed a theory of meaning which aims at reducing communicated meaning to speaker's meaning. He distinguished communicated meaning from *natural* meaning of a sign, and introduced the notion of *non-natural* meaning, or meaning$_{nn}$, for it:

> [$S$] means$_{nn}$ something by $x$' is roughly equivalent to '[$S$] uttered $x$ with the intention of inducing a belief by means of the recognition of this intention. (Grice, 1957, p. 385)

Speaker's meaning can be divided into explicit and implicit content. For Grice, the explicit content of an utterance is *what is said*, and the implicit content are the implicatures (Grice, 1989, Ch. 2).

An utterance not only has propositional content, it also has an illocutionary force (Searle, 1969). Speech acts are commonly assumed to be intentional acts. For the successful performance of assertive, directive, and commissive speech acts, it is necessary to communicate the act to an addressee. Hence, it is justified to call the whole speech act performed by an utterance *the speaker's meaning* of the utterance. In analogy to propositional content, we may distinguish between the explicit illocutionary force and the implicit illocutionary force of an utterance. We identify the former with the *direct*, and the latter with the *indirect* speech act (Searle, 1975).

The clause in Grice's definition of meaning$_{nn}$ which asks for the *recognition* of the speaker's intention implies explicit reasoning about the speaker's state of mind on the hearer's side when interpreting an utterance. In particular, it implies that the hearer first must infer the speaker's intention before he can infer the meaning of an utterance. For the following example from (Grice, 1989), this means that the implicature $R$ can only be inferred after inferring that the speaker intends to recommend that the hearer should look for petrol at the garage round the corner:

**(1)** $H$ is standing by an obviously immobilised car and is approached by $S$; the following exchange takes place:
$H$: I am out of petrol.
$S$: There is a garage round the corner. ($G$)
$+>$ The garage is open. ($R$)

The direct speech act type of answer $G$ is that of an assertive. In the given context, it acquires the indirect speech act type of a directive. Hence, we are interested in the question whether the recognition of the speaker's intentions is a prerequisite of recognising the directive speech act, or whether it follows it. We present a detailed game theoretic model of this example and study the order in which the inferences are drawn. This model shows that the *explicit* inferences about the speaker's beliefs about the garage follows *after* successfully inferring the relevance implicature $R$. Hence, the order of inferences is the inverse of that predicted by Grice.

This study of the role of speaker intentions serves as a model for answering the wider question of how much reasoning about each other is necessary for normal conversation. By *normal* conversation, we mean here conversation for which

it is common knowledge that the interlocutors are cooperative and have only true beliefs. We address this question in the framework of *iterated best response* models (Jäger and Ebert, 2009; Franke, 2009). In these models, the steps of reasoning about each other are made explicit. They start out with speaker and hearer strategies which are only constrained by semantic rules, and then step by step improve them by taking into account the other interlocutor's strategy. In this framework, it is possible to address the question of how much iteration of these improvement steps is necessary for reaching a stable equilibrium. We show that the *optimal answer* model (Benz and van Rooij, 2007) allows for the least number of steps. We take this as strong support for this model.

The next section introduces the relevant aspects of Grice theory of non–natural meaning. We then consider our problem within the framework of iterated best response models. Afterwards, we introduce the Optimal–Answer model for finding optimal answers and their implicatures. Finally, we study the role of the recognition of speaker's intentions in this model and discuss the general significance of our findings. We claim that, for a certain class of *normal* utterance situation, communication does *not* depend on the explicit recognition of the speaker's intentions or state of mind.

# 2 Meaning and conventions

Intuitively, speaker's meaning is the content the speaker intends to express and communicate by an utterance. In (Grice, 1957), he set out to clarify this concept of *speaker's meaning* by distinguishing it from *natural meaning*. Natural meaning is the information which can be carried by an event or object independently of the beliefs and intentions of any person who may use this event or object for the purposes of communication. Grice used the following example for illustrating the concept of natural meaning:

**(2)** **a)** Those spots mean measles.

**b)** Those spots didn't mean anything to me, but to the doctor they meant measles.

In both sentences, the word *meaning* refers to natural meaning. The spots carry the information that the patient is infected with measles independently of any person using the spots for communicating that he is infected with measles.

The counter–part to natural meaning is *non–natural* meaning, or *speaker's meaning*. The following definition from (Grice, 1989, p. 92) is more precise about the role of *causality*. Grice defined:

**(3)** "[$S$] meant something by uttering $x$" is true, iff for some audience $A$, [$S$] uttered $x$ intending:

**a)** *A* to produce a particular response *r*

**b)** *A* to think (recognize) that [*S*] intends (3a)

**c)** *A* to fulfil (3a) on the basis of his fulfilment of (3b).

Grice wants the clause *on the basis of* to be understood in the sense that the addressee's thinking that *S* intends him to respond with *r* is at least part of his reason to produce *r*, and not merely a cause for his producing *r*. To see this, let us replace (3c) by (3c') which merely asks for a causal relation between recognising *S*'s intention and the fulfilment of *r*:

**c')** *A* to fulfil (3a) as a result of his fulfilment of (3b).

This condition would entail that a speaker *S* would $mean_{nn}$ something by doing something *x* with the intended effect of (Grice, 1989, p. 92):

**(4)** **a)** *A* to be amused

**b)** *A* to think that [*S*] intended him to be amused

**c)** *A* to be amused (at least partly) as a result of his thinking that [*S*] intended him to be amused.

Clearly, to cause somebody to be amused by making him recognises that one tries to make him amused is not a case of communication. Therefore, Grice added the condition that the recognition of the intention must be a reason and not merely a cause. It is not exactly clear what *reason* means here; but we can assume that it is some kind of argument from which the addressee can infer that it is rational for him to produce the intended response *r*. Hence, it seems that clause (3c) implies that the recognition of the speaker's intention is a prerequisite for interpreting his utterance.[1]

There is a quite substantial amount of literature on Grice' theory of $meaning_{nn}$, see (Avramides, 1997) for a short overview. The discussion of examples and counter–examples led to various revisions and refinements, one of the most important concerned the role of *common knowledge* (Schiffer, 1972). Another important revision concerns the role of linguistic conventions. We here follow Searle, who pointed out that for a signalling act to count as linguistic communication it is necessary that the signal carries it's meaning in virtue of conventional rules. In

[1]This interpretation of clause (3c) is not the only possible interpretation. The clause could also be satisfied if the hearer first arrives at the interpretation of the utterance, from which he then recognises the speaker's intention, which in turn provides the essential *motivation* for producing the appropriate response *r*, i.e., for example, for believing in the speaker's assertion, fulfilling his request, or accepting his promise. This alternative interpretation has no effect on our analysis as we will argue that in *normal* utterance situations the motivation is already provided by the conventional meaning of the utterance.

(Searle, 1969, p. 49f), he proposes the following revised Gricean analysis of the *literal* meaning of an utterance:

**(5)** *S* utters sentence *T* and means it (i.e., means literally what he says) = *S* utters *T* and

**a)** *S* intends (*i*-I) the utterance *U* of *T* to produce in *H* the knowledge (recognition, awareness) that the state of affairs specified by (certain of) the rules of *T* obtain. (Call this effect the illocutionrary effect, *IE*).

**b)** *S* intends *U* to produce *IE* by means of the recognition of *i*-I.

**c)** *S* intends that *i*-I will be recognized in virtue of (by means of) *H*'s knowledge of (certain of) the rules governing (the elements of) *T*.

Both clause (5a) and clause (5c) contain a reference to *rules* which govern the interpretation of the sentence *T*. Searle explains these conditions using the example of greeting by uttering "*Hallo*". To *mean* this act of greeting to be a greeting involves "*(a) intending to get the hearer to recognize that he is greeted, (b) intending to get him to recognize that he is being greeted by means of getting him to recognize one's intention to greet him, (c) intending to get him to recognize one's intention to greet him in virtue of his knowledge of the meaning of the sentence 'Hallo'*" (Searle, 1969, p. 49). We see that this definition of (non–natural) meaning itself makes reference to some form of non–natural meaning. The type of *meaning* referred to in the (*c*) clause is *conventional* meaning. Searle, in contrast to Grice, does not try to reduce semantic meaning to speaker's meaning. Instead, he assumes that semantic meaning is determined by convention. So, we may ask, why is the recognition of the speaker's intention necessary if the literal meaning is conveyed by a convention? For Searle, this is necessary in order to turn the timeless conventional meaning of a sentence into the actual meaning communicated by the specific utterance event.

A crucial step in the investigation of the role of speaker's intentions for the interpretation of utterances is the explication of the notion of *conventions*. In this respect, we follow David Lewis (2002) who considered communication as a coordination problem which he described by a game theoretic model. For this purpose, he introduced *signalling* games, and defined conventional meaning using the equilibria of these games. In the simplest case, a signalling game consists of a set of speaker types $\Theta$ which represent the speaker's possible information states, a set of signals $\mathscr{F}$, and a set of hearer actions $\mathscr{A}$. The speaker knows his type $\theta \in \Theta$, which is not known to the hearer. In the simplest case, the speaker just wants to communicate his private type; hence, in this case, the hearer's action set $\mathscr{A}$ can be identified with $\Theta$. Apart from the speaker's type, the structure of the game is common knowledge. The game is sequential, i.e. first the speaker chooses

a signal $F \in \mathscr{F}$, and then, after receiving $F$, the hearer tries to guess the speaker's type $\theta$. The behaviour of speaker and hearer is represented by their *strategies*, i.e. in case of the speaker by a function $S : \Theta \to \mathscr{F}$ which maps each type $\theta$ to a signal $F$, and in case of the hearer by a function $H : \mathscr{F} \to \Theta$ which maps each signal $F$ back to a type $\theta' \in \Theta$. If the speaker's type is $\theta$, communication is successful if $H(S(\theta)) = \theta$. It is guaranteed to be always successful if $H \circ S = \mathrm{id}$. In the latter case, the pair $(S,H)$ is a *signalling convention*,[2] and the *conventional meaning* of a signal $F$ can be identified with either $S^{-1}(F)$ or $H(F)$. If $(S,H)$ is a signalling convention, then $(S,H)$ is a *Nash* equilibrium; i.e. there is no speaker strategy $S'$ such that the speaker would prefer playing $S'$ against $H$, and there is no hearer strategy $H'$ such that the hearer would prefer playing $H'$ against $S$.

The signals $F \in \mathscr{F}$ have no predefined meaning. They acquire meaning only in virtue of the signalling convention. This explains, among other things, why the choice of signals is arbitrary. If the interlocutors had agreed on another signalling convention $(S',H')$ for which there exist $\theta \neq \theta'$ and $F \neq F'$ such that $S'(\theta) = F' = S(\theta')$ and $S'(\theta') = F = S(\theta)$, then $F$ would mean $\theta$ for signalling convention $(S,H)$, and $\theta'$ for $(S',H')$.

In general, a signalling game may be much more complex. For example, the strategies may be probabilistic, the speaker's type may be a probability distribution over a set of possible worlds, and the set of hearer actions may be any set of actions. For example, Lewis (2002, p. 144) calls a signal $F$ *imperative* if $H(F)$ is an action in the everyday sense.

Lewis (2002, Sec. IV.5) shows that signalling $F$ with the intention to induce the hearer to believe or do $H(F)$ satisfies all of Grice's (1957) conditions for *meaning*$_{nn}$. This is a consequence of Lewis's (2002, p. 76) condition that for a strategy pair $(S,H)$ to count as a *convention* of a population it is necessary that it is *common knowledge* in the population. Roughly, we can reconstruct the speaker's reasoning using the above notation as follows: Assume that the speaker has private type $\theta$; as $(S,H)$ is a signalling convention, it follows that for speaker and hearer the outcome of $H \circ S(\theta)$ is the most desirable outcome; in this situation the speaker produces signal $S(\theta) = F$; then it follows that the speaker desires that his utterance will induce (a) the hearer to produce a particular response, namely $H(F) = a$; as $(S,H)$ is common knowledge, it follows that (b) the hearer knows that $S$ desires that (a); and (c) as the hearer knows that the speaker follows strategy $S$ and that the speaker utters $F$ because he wants to induce the hearer to do $a$, the hearer will actually do $H(F) = a$. Hence, Grice' conditions for meaning$_{nn}$ as stated in (3) are satisfied.

[2]This is, of course, a gross simplification of Lewis definition of *convention* (2002, p. 58 & p. 76) and *signalling convention* (2002, p. 135).

Let us compare Lewis's account with Searle's. Searle assumed that convention determines the timeless meaning of a sentence. In order to make the timeless meaning the actual meaning of its utterance, the Gricean conditions on the recognition of the speaker's intentions must in addition be satisfied. For Lewis, a convention is a regularity of behaviour described by a strategy pair. Hence, a sentence $F$ has a conventional meaning $M$ because it is used by a community to communicate $M$. There is no need to fill a gap between timeless meaning and occasional utterance meaning. Once the speaker signals $S(\theta) = F$ in situation $\theta$, the hearer will interpret it by $H(F) = M$. There is, of course, a caveat. Both Searle and Lewis only account for semantic meaning, i.e. for *what is said* in Grice's terms. They don't account for *non–conventional*, i.e. conversational, implicatures. According to Searle (1975), the indirect speech acts can be explained as implicatures which result from reasoning which takes into account conversational maxims, the direct speech act, and other contextual information. Both indirect speech acts and conversational implicatures are generally subsumed under meaning$_{nn}$. It seems to follow by definition that the recognition of implicatures involves reasoning about the speaker's state and his intentions, as the following quote from (Grice, 1989, p. 86) shows:

> "... what is implicated is what is required that one assume a speaker to think in order to preserve the assumption that he is observing the Cooperative Principle (and perhaps some conversational maxims as well), ..."

As Lewis remarks (2002, p. 155), it is actually not necessary to go through all the practical reasoning necessary to establish the Gricean conditions in order to perform an intentional act of communication. He claims that an act can be intentional without deliberation. Hence, the reasoning which we have seen before is a purely hypothetical reasoning. This leads us to the following two questions. First, how much explicit reasoning about each other is necessary for communicating *what is said*? And second, how much explicit reasoning about each other is necessary for communicating *what is implicated*?

# 3 Interpretation and reasoning about each other

For answering the question as to how much reasoning about each other is necessary for communicating what is said, we again consider Lewis's signalling games and signalling conventions. The assumption was that the game structure and the conventions are common knowledge. It is not necessary to assume that this common knowledge is represented and reasoned about consciously. All that is necessary is that the hearer plays his part of the game exactly iff the speaker plays the

other part. To see this, let $P$ be a given population of interlocutors each of whose members can play the role of the speaker or of the hearer. Let's assume that each interlocutor adopts strategy $S$ as a speaker and strategy $H$ as a hearer. Then let's further assume that two players are drawn at random and each player is assigned one of the roles without knowing the identity of the other player. Then, for all the speaker's epistemically possible speaker–hearer–pairs it is necessarily the case that the speaker follows $S$ and the hearer $H$; furthermore in all the speaker's epistemically possible pairs it is necessarily the case that in all the hearer's epistemically possible pairs the speaker follows $S$ and the hearer $H$; then it is also the case that in all the speaker's epistemically possible pairs it is necessarily the case that in all the hearer's epistemically possible pairs it is necessarily the case that in all the speaker's epistemically possible pairs the speaker follows $S$ and the hearer $H$, etc. Hence, simply from the fact that speaker and hearer are members of $P$ and the fact that they are fully cooperative, they can infer that the signalling convention $(S,H)$ is common knowledge. They both are fully justified in following $(S,H)$ without any deliberation about each other. The population $P$ is the population sharing the same language, and cooperativity can be thought of as a normality assumption which does not need to be consciously justified for each utterance situation. Hence, for communicating *what is said*, the hearer may just follow his interpretation strategy $H$ without deliberation about the speaker's intentions as long as there is no information that the speaker is non–cooperative or doesn't speak the same language. There is no need to make the practical reasoning which justifies this behaviour explicit.

From a more general perspective, we can look at a signalling convention, once it has been established in a population, as a natural regularity upon which the interlocutors can rely without further deliberation; in the very same way as car drivers don't need to reason about the reasoning of other car drivers when interpreting traffic lights. They stop if the lights show red, and drive if they are green. If everyone in the population adheres to this convention, no reasoning about each other is necessary for establishing common information that no car crashes can occur.

The situation is different if the information provided by the speaker has to be used to solve a decision problem as that of finding petrol in the Out–of–Petrol example (1), or if the interpretation of particularised implicatures or indirect speech acts are concerned. In these cases, the hearer, in general, cannot infer the optimal choice, the implicature, or the indirect illocutionary force simply from a signalling convention. Hence, they may involve explicit reasoning about the speaker's intentions. This is the problem which we investigate in the remainder of the paper.

Let us assume that there is a signalling convention which defines the *semantic* meaning of signals; i.e. the signals have a predefined meaning which restricts their use. The signalling convention also determines the hearer's interpretation

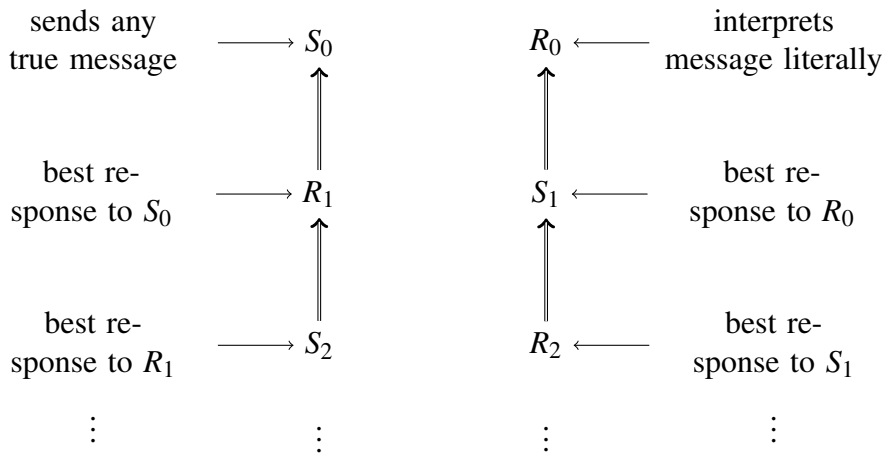


Figure 1: Schema of the IBR-sequence (Franke, 2009, p. 57).

of signals. We are interested in situations in which the hearer has to choose between several actions based on the information received from the speaker. We assume that cooperativity is assumed as a default and that the hearer has no reason to believe the opposite. Starting out from this situation, we are interested in the necessary amount of reasoning about each other for making sure that communication is successful. This can be made more precise in the framework of *iterated best response* (IBR) models (Jäger and Ebert, 2009; Franke, 2009).[3] IBR models explicate the reasoning about each other by an iterated process. In each step of this process, one of the two interlocutors chooses a best response strategy to the strategy which he assumes the other interlocutor has chosen in the previous step. There are two possible strategies from which the IBR process can start: the process can either start with a speaker strategy or with a hearer strategy. Accordingly, the model consists of two separate lines of reasoning. These two lines are shown in Figure 1.

In the IBR models worked out by (Jäger and Ebert, 2009; Franke, 2009), the $S_i$ and $R_i$ are in fact *sets* of strategies. In (Franke, 2009), $S_0$ is the set of all speaker strategies for which the speaker arbitrarily chooses a signal which he believes to be true. Hence, the $S_0$–speakers do not take the hearer's strategy into account. The hearer chooses an action after receiving the speaker's signal. Receiving it, he learns the semantic content of it. $R_0$ is the set of all hearer strategies for which the hearer only takes the semantic meaning of signals into account. Hence, $R_0$–hearers do not reason about the speaker. This means that on the 0–level it suffices

[3]The following sketch of the IBR model is a simplified version of (Franke, 2009). For more details, motivation, and differences between the models, we refer to the original papers.

to know the shared utilities and the speaker's and hearer's (subjective) probabilities about the state of affairs for defining $S_0$ and $R_0$. In step $n+1$ of the IBR process, each interlocutor $I$ assumes that the other interlocutor $J$ adopts a certain strategy from $J$'s strategy set defined in the $n$th step. Together with $I$'s expectations about the state of affairs, this defines $I$'s new set of best response strategies. This means, e.g., that, in the first step from $S_0$ to $R_1$, the hearer assumes that the speaker adopted some $S_0$ strategy, which arbitrarily chooses a sentence which the speaker believes to be true. After receiving a signal $F$, the hearer chooses an act whith the highest expected utility given the fact that the speaker sent $F$. $R_1$ is then the set of all hearer strategies which are best responses to some $S \in S_0$. Similarly, in the first iteration step from $R_0$ to $S_1$, the speaker assumes that the hearer follows some strategy in $R_0$. The speaker, as a response, chooses signals which lead the hearer to choose such actions which will have the highest expected utility as seen from the speaker's perspective. This defines the set $S_1$. This process can be iterated. IBR models then look for pairs of strategy sets $(S^*, H^*)$ which eventually become *stable*.[4]

What is the significance of this model in the context of our discussion of the role of speaker intentions in communication? A hearer who follows an $R_0$ strategy does not take the speaker into account, hence the speaker's intentions don't play a role for his interpretations. A speaker following a $S_0$ strategy doesn't take the hearer into account, hence no *intention* can be attributed to him. As a $R_1$ hearers only considers $S_0$ speakers, it follows that also for them speaker intentions don't play a role. The $S_1$ and $S_2$ speaker take into account how the hearer will react to their strategies, and they choose their signals such that the expected outcome of the hearer's choice is in their mutual best interest. Hence, we can call their signalling intentional acts. This entails that if it is necessary to move to the $R_2$ or $R_3$ level for reaching a stable equilibrium, then it follows that the hearer has to consider the speaker's intentions explicitly. If it is possible to reach a stable equilibrium without moving beyond level $R_0$ or $R_1$, then the recognition of speaker intentions doesn't play a role in interpretation. Hence, the answer to our second question, that about how much explicit reasoning about each other is necessary for communicating *what is implicated*, will follow once we know how many steps we have to move upwards in the IBR model for reaching an equilibrium.

For answering the latter question, we can consider the two lines of the IBR model separately as strategy sets occurring in one line have no influence on the strategy sets in the other line. Let us consider the line starting with the speaker strategies in $S_0$. The hearers set of best responses $R_1$ will in general be different from $R_0$ as the fact that a signal was sent may carry information in addition to the

[4]Stability is defined by a *looping* condition for the strategy sets $S^*$ and $R^*$. For details, see (Franke, 2009, p. 58).

semantic meaning of the signal. As the strategies in $S_0$ randomly produced true signals, $S_2$, the speaker's best responses to $R_1$, will in general be different from $S_0$. Hence, a stable state cannot be reached before $S_2$ is reached. The earliest stage at which the hearer can see that he has reached a stable state is therefore the stage in which he calculates $R_3$; and the earliest stage at which the speaker can see that he has reached a stable state is, accordingly, the stage in which he calculates $S_4$. Hence, for the line starting with $S_0$, for knowing that a stable state is reached, the hearer must at least consider the speaker's best response to his best response to the speaker's random strategy; and the speaker has at least to consider the hearer's best responses to the speaker's best responses to the hearer's best responses to the speaker's random strategies in $S_0$. Let us now turn to the line of the IBR model starting with $R_0$. The earliest stage at which the hearer can see that he has reached a stable state is the stage in which he calculates $R_2$; and the earliest stage at which the speaker can see that he has reached a stable state is, accordingly, the stage in which he calculates $S_3$. Hence, for the line starting with $R_0$, the hearer must at least consider the speaker's best response to his basic strategies in $R_0$, and the speaker has at least to consider the hearer's best responses to the speaker's best responses to the hearer's basic strategies. If we take the IBR model serious as a cognitive model, then these reasoning steps must be a cognitive reality.

To sum up, our first investigation of the IBR model seems to show that the hearer must take into account the speaker's best response to a hearer strategy at least once. Hence, the shortest possible path to a stable strategy is the $R_0$–$S_1$–$R_2$–$S_3$–path. This supports the Gricean (1957) view: If $S_3 = S_1$, the speaker following an $S_3$–strategy utters something with the intention to induce the hearer to an optimal act (as $S_1$–strategy), and the recognition of this intention (which takes place in the $R_2$–step) is a necessary condition for the hearer to arrive at the desired conclusion. In particular, the recognition of the speaker's intention is not only a *cause*, it is a *reason* for the hearer to follow his $R_2$–strategy.

But this does not jet show that the Gricean conditions must be satisfied. There may be other methods for finding equilibria which avoid the necessity for the hearer to reason about the speaker's intentions. If a method should be simpler or shorter than the method provided by the IBR model, then it has to avoid some steps of reasoning about each other in the IBR–sequence. In this respect, the simplest method is backward induction. When applying backward induction, the hearer does never consider the speaker's strategy, and the speaker considers the hearer's strategy only once. This is the cognitively least demanding method for finding solutions. It can be shown (a) that the resulting strategy pair $(S,H)$ guarantees that for any possible utterance the signal *naturally means* that the hearer chooses a speaker optimal act, and (b) that $(S,H)$ is a Pareto Nash equilibrium (Benz, 2009). Hence $(S,H)$ is a stable strategy pair, and there is no need for further steps of reasoning about each other.

# 4 The Optimal–Answer model

In this section, we show that the coordination problem posed by communication can be solved with fewer steps of reasoning about each other than predicted by the IBR model. More precisely, we show that backward induction provides a solution which guarantees that speaker and hearer have reached a stable strategy pair without having to calculate *whether they have reached a stable state*. The model, which we call the *the Optimal–Answer* (OA) model, was introduced in (Benz, 2006). In this and the following section, we introduce this model before we come back to the role of speaker's intention in communication in Section 6.

The general features of the communicative situation are the same as those considered in the context of signalling games. We assume that the conversation is subordinated to a joint purpose which is defined by a decision problem of the hearer. This decision problem may be revealed by an implicit or explicit question by the hearer. Hence, we can call the speaker's message an *answer*. The OA model tells us which answer a rational language user will choose given the hearer's decision problem and his knowledge about the world. We call the basic models which represent the utterance situation as *support problems*. They consist of the hearer's decision problem and the speaker's expectations about the world. These expectations are represented by subjective probabilities. In (Benz, 2006, 2007), it was shown that, in general, it is not possible to define a reliable *relevance* measure such that the speaker may simply maximise the relevance of his answers for optimally supporting the hearer. When solving a support problem the speaker has to take the hearer's response to his choice of signal into account. Hence, in view of our previous discussion of IBR models, this shows that there is no reliable method of solving a support problem which involves fewer steps of reasoning about each other than backward induction. Support problems incorporate Grice' *Cooperative Principle*, his maxim of *Quality*, and a method for finding optimal strategies which replaces Grice' maxims of *Quantity* and *Relevance*. We ignore the maxim of *Manner*.

A decision problem consists of a set $\Omega$ of the possible states of the world, the decision maker's expectations about the world, a set of actions $\mathscr{A}$ he can choose from, and his preferences regarding their outcomes. We always assume that $\Omega$ is finite. We represent an agent's expectations about the world by a probability distribution over $\Omega$, i.e. a real valued function $P : \Omega \to \mathbb{R}$ with the following properties: (1) $P(v) \geq 0$ for all $v \in \Omega$ and (2) $\sum_{v\in\Omega} P(v) = 1$. For sets $F \subseteq \Omega$ it is $P(F) = \sum_{v\in F} P(v)$. The pair $(\Omega, P)$ is called a finite *probability space*. An agent's preferences regarding outcomes of actions are represented by a real valued function over world–action pairs. We collect these elements in the following structure:

**Definition 4.1** *A* decision problem *is a triple* $\langle(\Omega, P), \mathscr{A}, u\rangle$ *such that* $(\Omega, P)$ *is a*

*finite probability space, $\mathscr{A}$ a finite, non–empty set and $u : \Omega \times \mathscr{A} \to \mathbb{R}$ a function. $\mathscr{A}$ is called the* action *set, and its elements* actions*; $u$ is called a* payoff *or* utility *function.*

In the following, a decision problem $\langle(\Omega, P), \mathscr{A}, u\rangle$ represents the hearer's situation before receiving information from an answering expert. We will assume that this problem is common knowledge. How to find a solution to a decision problem? It is standard to assume that rational agents choose actions $a$ for which the expected utility $EU(a)$ is maximal:

$$EU(a) = \sum_{v \in \Omega} P(v) \times u(v, a). \tag{4.1}$$

The expected utility of actions may change if the decision maker learns new information. To determine this change of expected utility, we first have to know how learning new information affects the hearer's beliefs. In probability theory the result of learning a proposition $F$ is modelled by *conditional probabilities*. Let $A$ be any proposition and $F$ the newly learned proposition. Then, the probability of $A$ *given* $F$, written $P(A|F)$, is defined as

$$P(A|F) := P(A \cap F)/P(F) \text{ for } P(F) \neq 0. \tag{4.2}$$

In terms of this conditional probability function, the *expected utility after learning* $F$ is defined as

$$EU(a|F) = \sum_{v \in \Omega} P(v|F) \times u(v, a). \tag{4.3}$$

$H$ will choose the action which maximises his expected utilities after learning $F$, i.e. he will only choose actions $a$ for which $EU(a|F)$ is maximal. We assume that $H$'s decision does not depend on what he believes that the answering speaker believes. We denote the set of actions with maximal expected utility by $\mathscr{B}(F)$, i.e.

$$\mathscr{B}(F) := \{a \in \mathscr{A} \mid \forall b \in \mathscr{A}\ EU_H(b|F) \leq EU_H(a|F)\}. \tag{4.4}$$

The decision problem represents the hearer's situation. In order to get a model of the questioning and answering situation, we have to add a representation of the answering speaker's information state. We identify it with a (subjective) probability distribution $P_S$ that represents his expectations about the world. We make a number of assumptions which define a certain class of *normal* utterance situations. First, we assume that the hearer's expectations are common knowledge. Second, we assume that there exists a common prior from which both the speaker's and the hearer's information state can be derived by a Bayesian update. This entails that the speakers and the hearer's expectations cannot contradict each other. Third,

we assume that the speaker does not directly choose propositions but linguistic *forms* or *signals* which have a predefined semantics. This leads to the following definition of *interpreted* support problems:

**Definition 4.2** *A tuple* $\sigma = \langle \Omega, P_S, P_H, \mathscr{F}, \mathscr{A}, u, [\![ . ]\!] \rangle$ *is an* interpreted support problem *if: (1)* $(\Omega, P_S)$ *is a finite probability space and* $\langle (\Omega, P_H), \mathscr{A}, u \rangle$ *a decision problem; (2) there exists a probability distribution P on* $\Omega$*, and sets* $K_S \subseteq K_H \subseteq \Omega$ *for which* $P_S(X) = P(X|K_S)$ *and* $P_H(X) = P(X|K_H)$*; (3)* $[\![ . ]\!] : \mathscr{F} \to \mathscr{P}(\Omega)$ *is an interpretation function for the elements* $F \in \mathscr{F}$*; and (4)* $u : \Omega \times \mathscr{A} \to \mathbb{R}$ *is a utility measure. We assume in addition that* $\forall X \subseteq \Omega \, \exists F \in \mathscr{F} \, [\![F]\!] = X$.

The second condition says that $P_S$ and $P_H$ are derived from a common prior $P$ by a Bayesian update. It entails that $\forall X \subseteq \Omega \; P_S(X) = P_H(X|K_S)$. This condition allows us to identify the *common ground* in conversation with the addressee's expectations about the domain $\Omega$, i.e. with $P_H$. The speaker knows the addressee's information state and is at least as well informed about $\Omega$. Hence, the assumption is a probabilistic equivalent to the assumption about common ground that implicitly underlies dynamic semantics (Groenendijk and Stockhof, 1991). The condition furthermore implies that the speaker's beliefs cannot contradict the hearer's expectations, i.e. for $X \subseteq \Omega$: $P_S(X) = 1 \Rightarrow P_H(X) > 0$. In order to simplify notation, we will often write $F$ instead of $[\![F]\!]$. Hence, $F$ may denote a proposition or a linguistic form, depending on context.

Our next goal is to introduce a principle for solving support problems, i.e. for finding the speaker's and hearer's strategies which lead to optimal outcomes. The speaker $S$'s task is to provide information that is optimally suited to support $H$ in his decision problem. We assume that $S$ is fully cooperative and wants to maximise $H$'s final success; i.e. $S$'s payoff, is identical with $H$'s. This is our representation of Grice's *Cooperative Principle*. $S$ has to choose an answer that induces $H$ to choose an action that maximises their common payoff. In general, there may exist several equally optimal actions $a \in \mathscr{B}(F)$ which $H$ may choose. Hence, the expected utility of an answer depends on the probability with which $H$ will choose the different actions. We can assume that this probability is given by a probability measure $h(.|F)$ on $\mathscr{A}$. Then, the expected utility of an answer $F$ is defined by:

$$EU_S(F) := \sum_{a \in \mathscr{B}(F)} h(a|F) \times EU_S(a). \tag{4.5}$$

We add here a further Gricean maxim, the *Maxim of Quality*. We call an answer $F$ *admissible* if $P_S(F) = 1$. The Maxim of Quality is equivalent to the assumption that the speaker is restricted to admissible answers; i.e. to answers

which he believes to be *true*. The set of admissible answers is defined as:

$$\mathrm{Adm}_\sigma := \{F \subseteq \Omega \,|\, P_S(F) = 1\} \tag{4.6}$$

Hence, the set of optimal answers in $\sigma$ is given by:

$$\mathrm{Op}_\sigma := \{F \in \mathrm{Adm}_\sigma \,|\, \forall B \in \mathrm{Adm}_\sigma\, EU_S(B) \leq EU_S(F)\}. \tag{4.7}$$

$\mathrm{Op}_\sigma$ is the set of *optimal answers* for the support problem $\sigma$. The definition of support problems entails that all propositions $A \subseteq \Omega$ can be expressed. We can think of $\mathrm{Op}_\sigma$ as a subset of $\mathscr{P}(\Omega)$ or as a subset of $\mathscr{F}$.

As mentioned before, the *behaviour* of interlocutors can be represented by *strategies*. In Section 2, we have seen strategies which choose for each information state a unique action. But uniqueness cannot always be assumed to hold for the choices of speakers and hearers. A *mixed* strategy is a strategy which chooses actions with certain probabilities. We define a (mixed) strategy pair for an interpreted support problem $\sigma$ to be a pair $(S,H)$ such that $S$ is a probability distribution over $\mathscr{F}$ and $H(.|F)$ a probability distribution over $\mathscr{A}$. We may call a strategy pair $(S,H)$ a *solution* to $\sigma$ iff $H(.|F)$ is a probability distribution over $\mathscr{B}(F)$, and $S$ a probability distribution over $\mathrm{Op}_\sigma$.

In general, the solution to a support problem is not uniquely defined. Therefore, we introduce the notion of the *canonical* solution.

**Definition 4.3** *Let $\sigma$ be a given interpreted support problem. The* canonical solution *to $\sigma$ is a pair $(S,H)$ of mixed strategies for which: (1) $S(F) = |\mathrm{Op}_\sigma|^{-1}$ if $F \in \mathrm{Op}_\sigma$, and $S(F) = 0$ else; (2) $H(a|F) = |\mathscr{B}(F)|^{-1}$ if $a \in \mathscr{B}(F)$, and $H(a|F) = 0$ else.*

We write $S(\,.\,|\sigma)$ if $S$ is a function that maps each $\sigma$ of a set $\mathscr{S}$ of support problems to the speaker's canonical strategy. From now on, we will always assume that speaker and hearer follow the canonical solution.

It can be shown (Benz, 2009) that the canonical solution $(S,H)$ (weakly) Pareto dominates all other strategy pairs. Hence, once the canonical solution is reached by backward induction, no further reasoning about each other is necessary.

# 5 Implicatures of optimal answers

An implicature of an utterance is a proposition which is implied by the assumption that the speaker is cooperative and observes the conversational maxims. More precisely, Grice linked implicatures to what the hearer learns from the utterance about

the speaker's knowledge. The speaker's canonical solution maps his possible information states to utterances. Hence, the hearer can use this strategy to calculate what the speaker must have known when making his utterance. As the canonical solution is a solution, it also incorporates the information that the speaker is cooperative and follows the maxims.

We treat all implicatures as particularised implicatures, i.e. as implicatures that follow immediately from the maxims and the particular circumstances of the utterance context. The answering expert knows a proposition $H$ in a situation $\sigma$ iff $P_E^\sigma(H) = 1$. Hence, if the inquirer wants to know what the speaker knew when answering that $A$, he can check all his epistemically possible support problems for what the speaker believes in them. If $\sigma$ is the support problem which represents the actual answering situation, then all support problems $\hat{\sigma}$ with the same decision problem are indistinguishable for the inquirer. Hence, the inquirer knows that the speaker believed that $R$ when making his utterance $A$, iff the speaker believes that $R$ in all indiscernible support problems in which $A$ is an optimal answer. This leads to the following definition:

**Definition 5.1 (Implicature)** *Let $\mathscr{S}$ be a given set of support problems with joint decision problem $\langle(\Omega, P_H), \mathscr{A}, u\rangle$. Let $A, R \subseteq \Omega$ be two propositions with $A \in \mathrm{Op}_\sigma$ for some $\sigma \in \mathscr{S}$. Then we set:*

$$A +\!> R \Leftrightarrow \forall \sigma \in \mathscr{S}\,(A \in \mathrm{Op}_\sigma \rightarrow P_E^\sigma(R) = 1), \tag{5.8}$$

*If $A +\!> R$, we say that the* utterance of $A$ implicates that $R$.

We are now interested in cases in which the speaker is a real expert. If he is an expert, then we can show that there is a very simple criterion for calculating implicatures. We can call the speaker an expert if he knows the actual world; but we will see that a weaker condition is sufficient for our purposes. To make precise what we mean by expert, we introduce another important notion, the set $O(a)$ of all worlds in which an action $a$ is optimal:

$$O(a) := \{w \in \Omega \,|\, \forall b \in \mathscr{A}\; u(w,a) \geq u(w,b)\}. \tag{5.9}$$

We say that the answering person is an expert for a decision problem if there is an action which is an optimal action in all his epistemically possible worlds. We represent this information in $\mathscr{S}$:

**Definition 5.2 (Expert)** *Let $\mathscr{S}$ be a set of support problems with joint decision problem $\langle(\Omega, P_H), \mathscr{A}, u\rangle$. Then we call $E$ an* expert *in a support problem $\sigma$ if $\exists a \in \mathscr{A}\; P_E^\sigma(O(a)) = 1$. He is an expert in $\mathscr{S}$, if he is an expert in every $\sigma \in \mathscr{S}$.*

This leads us to the following criterion for implicatures:[5]

**Lemma 5.3** *Let $\mathscr{S}$ be a set of support problems with joint decision problem $\langle(\Omega,P_H),\mathscr{A},u\rangle$. Assume furthermore that $E$ is an expert for every $\sigma\in\mathscr{S}$ and that $\forall v\in\Omega\,\exists\sigma\in\mathscr{S}\,P_E^\sigma(v)=1$. Let $\sigma\in\mathscr{S}$ and $A,R\subseteq\Omega$ be two propositions with $A\in\mathrm{Op}_\sigma$. Then, with*

$$A^*:=\{v\in\Omega\,|\,P_I(v)>0\}\ \textit{and}\ A^+=\bigcap\{O(a)\,|\,a\in\mathscr{B}(A)\},$$

*it follows that $A+\!>R$ iff $A^*\cap A^+\subseteq R$.*

$A^*$ is the equivalent to the common ground updated with $A$. In the context of a support problem, we can interpret an answer $A$ as a *recommendation* to choose one of the action in $\mathscr{B}(A)$. We may say that the recommendation is *felicitous* only if all recommended actions are optimal. Hence, $A^+$ represents the information that follows from the felicity of the speech act of recommendation which is associated to the answer.

# 6 Speaker's intentions and implicatures

In this section, we take up and conclude our discussion of the role of the recognition of the speaker's intentions for inferring implicatures. In the previous two sections, we have seen a method for calculating optimal answers and inferring their implicatures based on backward induction. As mentioned before, it can be shown that the resulting strategy pair $(S,H)$ is a Pareto Nash equilibrium (Benz, 2009). This guarantees that $(S,H)$ is a stable strategy pair of the IBR sequence, while freeing the interlocutors from actually calculating that it is stable. Hence, in the OA model the equilibrium is reached after a $R_0$–$S_1$ reasoning sequence. This involves no reasoning about the speaker's intentions.

In the OA model, it is guaranteed to the hearer that he finds the action $a$ which is recommended by the speaker by just taking the semantic content of an utterance into account. Hence, he first recognises the intended perlocutionary effect. In contrast to calculating the perlocutionary effect and the optimal choice, it seems that the calculation of implicatures involves reasoning about the speaker's intentions. As the condition for calculating implicatures in (5.8) involves a universal quantification over all support problems $\sigma$, and the consideration of all speaker information states $P_S^\sigma$ for which the given answer is optimal, it seems that the hearer has to take the speaker's perspective into account and calculate for all possible speaker states his optimal answers. This means that the hearer has to calculate the

[5]For a proof see (Benz, 2009).

speaker's strategy $S$, and hence his intentions. In general, this cannot be avoided; but it can be avoided if the speaker is known to be an expert. For this special case, Lemma 5.3 provides a simple criterion. In particular, it avoids all reasoning about the speaker's intentions. As an example, we study the Out–of-Petrol example (Grice, 1989, p. 31):

**(6)** $H$: I am out of petrol.
$S$: There is a garage round the corner. ($G(d)$)
$+>$ The garage is open. ($R(d)$)

We can assume that B's assertion is an answer to the question "*Where can I buy petrol for my car*?" Let $d$ be the place of the garage. We distinguish four worlds $\{w_1, w_2, w_3, w_4\}$ and two actions $\{\text{go-to-d}, \text{search}\}$. Let $G(d)$ mean that $d$ is a petrol station, and $R(d)$ that it is some place open for customers. Let the worlds and utilities be defined as shown in the following table:

| $\Omega$ | $G(d)$ | $R(d)$ | go-to-d | search |
|---|---|---|---|---|
| $w_1$ | $+$ | $+$ | 1 | $\varepsilon$ |
| $w_2$ | $+$ | $-$ | 0 | $\varepsilon$ |
| $w_3$ | $-$ | $+$ | 0 | $\varepsilon$ |
| $w_4$ | $-$ | $-$ | 0 | $\varepsilon$ |

The answering expert knows that he is in $w_1$. We assume that $P_I$ and $\varepsilon$ are such that $EU_I(\text{go-to-d}|G(d)) > \varepsilon$, i.e. the inquirer thinks that the expected utility of going to that garage is higher than doing a random search in the town. Hence $\mathscr{B}(G(d)) = \{\text{go-to-d}\}$.

Semantics provides the meaning $G$ of the answer as well as its direct illocutionary assertive force. The dialogue situation is such that the answer is embedded in a decision problem of the hearer. As this is common knowledge by assumption, every assertion $A$ will automatically have the perlocutionary effect that it leads the hearer to choose an action from $\mathscr{B}(A)$. As $\mathscr{B}(G(d)) = \{\text{go-to-d}\}$, it follows that the answer $G$ must be interpreted as a recommendation to go to $d$. We see that $O(\text{go-to-d}) = \{w_1\} \subseteq R(d)$. Neither the calculation of $\mathscr{B}(G(d))$, nor of $O(\text{go-to-d})$ involves a consideration of the speaker's strategy $S$. By Lem. 5.3, it holds that $G(d) +> R(d)$ iff $G(d)^* \cap G(d)^+ \subseteq R$. $G(d)^+$ states the conditions under which the recommendation of going to $d$ is optimal. As $G(d)^+ = \bigcap\{O(a) \,|\, a \in \mathscr{B}(G(d))\} = O(\text{go-to-d}) = \{w_1\} \subseteq R(d)$, it follows that $G(d) +> R(d)$. Hence, we see that it is not necessary to know the speaker's intentions for calculating the implicature.

Finally, we have to consider the wider significance of our result. As we have seen, for Grice it was a defining property of meaning$_{nn}$ that the hearer has to recognise the speaker's intentions. Implicatures are generally assumed to be part

of meaning$_{nn}$. But, as we have seen in the Out–of–Petrol example, the hearer can infer a relevance implicature without explicitly calculating the speaker's intentions or possible information states. Hence, not all forms of non–natural communication involve an explicit recognition of the speaker's intentions. This is the main result we were seeking to find in this paper. We have also seen that this result depends on certain assumptions about the utterance situation. These assumptions involve that the speaker is fully cooperative, that he has full knowledge about the hearer's information state, and that he is an expert with respect to a decision problem which the hearer has to solve. These assumptions are strongly limiting the model, but none of them are unusual idealisations or at odds with standard assumptions about *normal* utterance situations. For example, in order to arrive at the standard scalar implicature *some* $+>$ *not all*, it is necessary to assume that the speaker is an expert as witnessed by an example like "*I believe that some of the students failed*". The assumption that the hearer's information state can be identified with the common ground, and hence is known to the speaker, is an assumption implicit in most frameworks of dynamic semantics, as mentioned before.

# 7 Conclusion

We set out discussing the validity of the claim that the recognition of the speaker's intention is a necessary part of successfully interpreting meaning$_{nn}$. As an example, we considered the explanation of the relevance implicature and the (indirect) directive speech acts in the Out–of–Petrol example within the optimal answer model. We have seen that neither the recognition of the perlocutionary effect, nor of the relevance implicature involves the recognition of the speaker's intentions. We take this as evidence that the Gricean conditions for meaning$_{nn}$ don't have to be consciously verified by the interlocutors.

# References

Avramides, A. (1997). Intention and Convention. In Hale, B. and Wright, C., editors, *A Companion to the Philosophy of Language*, pages 60–86. Blackwell Publishing, Oxford.

Benz, A. (2006). Utility and Relevance of Answers. In Benz, A., Jäger, G., and van Rooij, R., editors, *Game Theory and Pragmatics*, pages 195–214. Palgrave Macmillan, Basingstoke.

Benz, A. (2007). On Relevance Scale Approaches. In Puig-Waldmller, E., editor, *Proceedings of the Sinn und Bedeutung 11*, pages 91–105.

Benz, A. (2009). Outline of the Foundations for a Theory of Implicatures. In Benz, A. and Blutner, R., editors, *Papers on Pragmasemantics*, volume 51 of *ZAS Papers in Linguistics*, pages 153–201. Center for General Linguistics, Berlin.

Benz, A. and van Rooij, R. (2007). Optimal assertions and what they implicate: a uniform game theoretic approach. *Topoi - an International Review of Philosophy*, 27(1):63–78.

Franke, M. (2009). *Signal to Act: Game Theory in Pragmatics*. PhD thesis, Universiteit van Amsterdam. ILLC Dissertation Series DS-2009-11.

Grice, H. P. (1957). Meaning. *Philosophical Review*, 66:377–388.

Grice, H. P. (1989). *Studies in the Way of Words*. Harvard University Press, Cambridge MA.

Groenendijk, J. and Stockhof, M. (1991). Dynamic predicate logic. *Linguistics & Philosophy*, 14:39–100.

Jäger, G. and Ebert, C. (2009). Pragmatic Rationalizability. In Riester, A. and Solstad, T., editors, *Proceedings of Sinn und Bedeutung*, volume 13.

Lewis, D. (2002). *Convention*. Blackwell Publishers, Oxford. First published by Harvard University Press 1969.

Schiffer, S. (1972). *Meaning*. Clarendon Press, Oxford.

Searle, J. R. (1969). *Speech Acts*. Cambridge University Press, Cambridge.

Searle, J. R. (1975). Indirect Speech Acts. In Cole, P. and Morgan, J. L., editors, *Syntax and Semantics*, volume 3, pages 59–82. Academic Press, New York.